Registered Replication Report

ASSOCIATION FOR PSYCHOLOGICAL SCIENCE

Advances in Methods and Practices in Psychological Science
2018, Vol. 1(3) 299–317

DOI: 10.1177/2515245918781032
www.psychologicalscience.org/AMPPS

Registered Replication Report on Mazar, Amir, and Ariely (2008)

Bruno Verschuere*, Ewout H. Meijer*, Ariane Jim*, Katherine Hoogesteyn*, Robin Orthey*, Randy J. McCarthy*, John J. Skowronski*, Oguz A. Acar, Balazs Aczel, Bence E. Bakos, Fernando Barbosa, Ernest Baskin, Laurent Bègue, Gershon Ben-Shakhar, Angie R. Birt, Lisa Blatz, Steve D. Charman, Aline Claesen, Samuel L. Clay, Sean P. Coary, Jan Crusius, Jacqueline R. Evans, Noa Feldman, Fernando Ferreira-Santos, Matthias Gamer, Sara Gomes, Marta González-Iraizoz, Felix Holzmeister, Juergen Huber, Andrea Isoni, Ryan K. Jessup, Michael Kirchler, Nathalie klein Selle, Lina Koppel, Marton Kovacs, Tei Laine, Frank Lentz, David D. Loschelder, Elliot A. Ludvig, Monty L. Lynn, Scott D. Martin, Neil M. McLatchie, Mario Mechtel, Galit Nahari, Asil Ali Özdoğru, Rita Pasion, Charlotte R. Pennington, Arne Roets, Nir Rozmann, Irene Scopelliti, Eli Spiegelman, Kristina Suchotzki, Angela Sutan, Peter Szecsi, Gustav Tinghög, Jean-Christian Tisserand, Ulrich S. Tran, Alain Van Hiel, Wolf Vanpaemel, Daniel Västfjäll, Thomas Verliefde, Kévin Vezirian, Martin Voracek, Lara Warmelink, Katherine Wick, Bradford J. Wiggins, Keith Wylie, and Ezgi Yıldız

*Lead authors

Multilab direct replication of: Experiment 1 from Mazar, N., Amir, O., & Ariely, D. (2008). The dishonesty of honest people: A theory of self-concept maintenance. *Journal of Marketing Research*, *45*, 633–644. doi:10.1509/jmkr.45.6.633

Protocol vetted by: Nina Mazar, Dan Ariely, and On Amir

Abstract

The self-concept maintenance theory holds that many people will cheat in order to maximize self-profit, but only to the extent that they can do so while maintaining a positive self-concept. Mazar, Amir, and Ariely (2008, Experiment 1) gave participants an opportunity and incentive to cheat on a problem-solving task. Prior to that task, participants either recalled the Ten Commandments (a moral reminder) or recalled 10 books they had read in high school (a neutral task). Results were consistent with the self-concept maintenance theory. When given the opportunity to cheat, participants given the moral-reminder priming task reported solving 1.45 fewer matrices than did those given a neutral prime (Cohen's d = 0.48); moral reminders reduced cheating. Mazar et al.'s article is among the most cited in deception research, but their Experiment 1 has not been replicated directly. This Registered Replication Report describes the aggregated result of 25 direct replications (total N = 5,786), all of which followed the same preregistered protocol. In the primary meta-analysis (19 replications, total n = 4,674), participants who were given an opportunity

Corresponding Authors:
Bruno Verschuere, Department of Clinical Psychology, University of Amsterdam, Nieuwe Achtergracht 129B, 1018 VZ Amsterdam, The Netherlands
E-mail: b.j.verschuere@uva.nl
Ewout H. Meijer, Department of Clinical Psychological Science, Maastricht University, PO Box 616, 6200 MD Maastricht, The Netherlands
E-mail: eh.meijer@maastrichtuniversity.nl

to cheat reported solving 0.11 more matrices if they were given a moral reminder than if they were given a neutral reminder (95% confidence interval = [–0.09, 0.31]). This small effect was numerically in the opposite direction of the effect observed in the original study (Cohen's d = –0.04).

Keywords
cheating, morality, honesty, replication, Many Labs, open data, open materials, preregistered

Cheating is widespread and associated with substantial costs to society. As many as 60% of taxpayers evade taxes (Slemrod, 2007), and in 2011, global tax evasion was estimated to exceed US$3.1 trillion (The Tax Justice Network, 2011). The self-concept maintenance theory (Mazar, Amir, & Ariely, 2008) holds that people try to maximize self-profit while maintaining a positive self-concept regarding their honesty. This theory predicts that many people will cheat to benefit themselves, provided that they can preserve their positive self-concept. Often, this means that people will justify small amounts of cheating. For example, participants who were asked to privately roll a die and to report the outcome, and who would receive greater financial gain with higher rolls, reported an average roll of 3.63. That amount is above what a random roll of a die should produce on average (3.50), but far from the maximum possible value of 6 (Halevy, Shalvi, & Verschuere, 2014).

According to the self-concept maintenance theory, people should be less likely to cheat when they think about their own honesty than when they do not. In a well-known test of this prediction, Mazar et al. (2008, Experiment 1) gave participants (N = 229) an opportunity and incentive to cheat on a problem-solving task. Prior to that task, participants either recalled the Ten Commandments (a moral reminder) or recalled 10 books they had read in high school (a neutral task). The problem-solving task was embedded among other filler tasks in a large booklet, and it required participants to find numbers that add up exactly to 10 in each of a series of matrices (e.g., 3.81 and 6.19; see Fig. 1). After completing the task, half of the participants ripped the matrices sheet out of the booklet and wrote down the number of matrices they had solved on a separate scoring sheet in the booklet, thus having the opportunity to cheat. As a financial incentive for cheating, the task instructions stated that 2 randomly selected participants would receive $10 for each matrix they reported solving. Thus, the participants in this condition had both an incentive and an opportunity to cheat; they could receive payment, and the only record of the number of matrices they had solved was their own self-report. Participants who tried to recall books they had read prior to the matrix task claimed to have solved 4.22 matrices, whereas participants who had listed the Ten Commandments claimed to have solved 2.77 matrices. The other half of the participants were allocated to a control condition and did not tear out the matrices page, so there was no opportunity to cheat without being caught. In the control condition, participants primed with recalling books and those primed with recalling the Ten Commandments solved similar numbers of matrices (3.06 and 3.12, respectively). Together, this pattern of results suggests that people who had the opportunity to cheat after recalling the 10 books cheated, whereas those who had the opportunity to cheat after recalling the Ten Commandments did not cheat (their performance was lower by 1.45 matrices; Cohen's d = 0.48).

Mazar et al.'s (2008) article had been cited more than 1,600 times on Google Scholar as of April 2018, and it has helped inspire research on how religious and other moral primes affect honest behavior (for a review, see Rosenbaum, Billinger, & Stieglitz, 2014). The Ten Commandments study also has political implications: It was cited as a critical building block of self-concept maintenance theory in a set of policy recommendations made to President Obama as part of the REVISE model (Ayal, Gino, Barkan, & Ariely, 2015). Knowing the size and reliability of this effect is critical both for policymakers and for the self-concept maintenance theory. Yet the literature includes no direct replication attempts for this study.

This Registered Replication Report (RRR) project was designed to provide an accurate and precise estimate

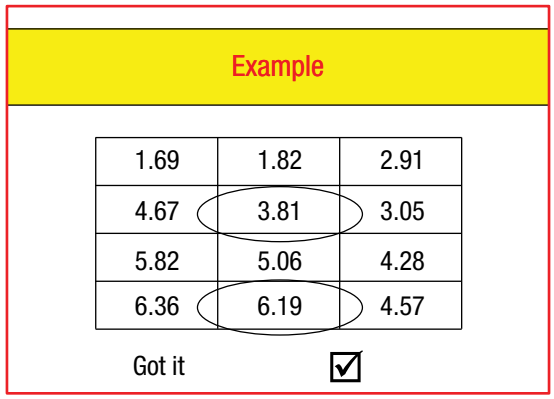

Fig. 1. Example matrix shown to participants in the instructions for the matrix task. Participants were told to search for the numbers that add up exactly to 10 (in this case, 3.81 and 6.19) and to mark the "Got it" box for each matrix they solved.

of the effect of Ten Commandments priming on cheating in the matrix task. The focus was on estimating the difference between the Ten Commandments priming condition and the 10-books priming condition in the number of matrices people reported solving. Given that the original study was conducted in the United States, and the vast majority of U.S. inhabitants identify themselves as Christians (Pew Research Center, 2015), we expected that there might be a different outcome in other cultures (Van Bavel, Mende-Siedlecki, Brady, & Reinero, 2016). Consequently, we examined the heterogeneity of the effect across laboratories and also included measures of religiousness to test the prediction that the effect of moral priming will be larger in laboratories whose participants hold stronger religious beliefs. These measures were administered after the primary tasks so that they would not influence the main outcome measure.

The RRR project was announced by the Association for Psychological Science and on social media, and laboratories were invited to apply to contribute. By the deadline of November 30, 2016, the Editor had received and approved 29 applications. Twenty-five out of these 29 contributing laboratories completed the study and provided data for the meta-analysis (see the appendix following the Discussion section for a list of the authors participating at each lab).[1] The study was administered as part of a larger task battery that included tasks for another RRR project reported in this issue (McCarthy et al., 2018), and all the contributing researchers are authors of both articles.

Disclosures

Preregistration

The approved protocol for the RRR was posted on the Open Science Framework (OSF) project page at https://osf.io/3bwx5/. Each laboratory preregistered their Editor-approved implementation of the official protocol on their individual project page, and those preregistrations are available by visiting the labs' project pages (linked from the Contributing Labs section at https://osf.io/hrju6/wiki/home/). Each laboratory team reported (on their project page) how they determined their sample size and documented all data exclusions. Any departures from the official protocol or the lab's preregistered implementation are documented in the Lab Implementation Appendix at https://osf.io/uskr8/ (also at http://journals.sagepub.com/doi/suppl/10.1177/2515245918781032). Drafts of the meta-analysis scripts were written in a data-blind manner, using simulated data. Those preregistered versions are posted at https://osf.io/jp45u/. The final scripts were updated to address minor formatting inconsistencies across labs, to improve the appearance of figures, and to add exploratory analyses. All changes from the data-blind scripts are noted in the final scripts posted at https://osf.io/mcvt7/.

Data, materials, and online resources

All materials are available at https://osf.io/rbejp/. All data and analyses are available at https://osf.io/mcvt7/wiki/home/. Supplementary online materials include the Lab Implementation Appendix, which documents the individual labs' contributions to the project (https://osf.io/uskr8/ and http://journals.sagepub.com/doi/suppl/10.1177/2515245918781032).

Reporting

We report how we determined our sample size, all data exclusions, all manipulations, and all measures in the study.

Ethical approval

Each laboratory obtained any necessary institutional-review-board or ethical approval from their home institution to accommodate differences in the requirements at different universities and in different countries.

Method

The protocol for this study was developed in consultation with the original authors, Nina Mazar, On Amir, and Dan Ariely, who provided their materials and feedback on key aspects of the design. The final protocol and materials were approved by the original authors and are publicly available at https://osf.io/vxz7q/. Note that participating labs were responsible for their own informed-consent forms and for translation of the materials (if testing was not conducted in American English or Dutch). When translations were required, laboratories were asked to translate the materials and then independently back-translate them to ensure accuracy of the translations. The Editor helped coordinate translations so that all laboratories testing in a given language used the same materials.

Testing took place in large classrooms. At least 50 participants were present simultaneously in each session, to ensure adequate anonymity, as in the original study, which was run in one single session (labs were asked to aim for 100 participants or more in each session). Each lab was required to collect usable data from at least 200 participants, 20% to 80% of whom were female. The tasks for this study were embedded in a

larger battery of tasks (stapled into a packet), all of which were completed using paper and pencil. In total, there were eight different versions of the packet (the four conditions for this study crossed with two conditions for the other replication study completed as part of the battery). Experimenters randomly shuffled the printed packets prior to each session to ensure random assignment of participants to conditions, and each packet included a cover page to mask the condition. The order of tasks in the packet was the same for all participants and is listed in Table 1. Completion of the whole test battery took approximately 45 min. Participants were informed that some of the tasks would require them to time themselves, and a stopwatch was projected on a screen at the front of the room for that purpose. After providing written informed consent, participants worked through the tasks in the booklet.

Design and procedure

Participants were required to be 18- to 25-year-old students. They were compensated with extra credit, module credit, course credit, or other nonmonetary rewards (e.g., free workshop attendance, movie tickets).

Table 1. List of Tasks in the Combined Procedure for the Two Registered Replication Reports (RRRs)

Task	Description	RRR
Demographics and informed consent	Participants provided their age, sex, and major and gave written informed consent.	Both
Sentence descrambling (hostility priming) (Srull & Wyer, 1979, Experiment 1)	For each of 30 groups of four words, participants marked the three words that would make a complete sentence (e.g., "child the question watch"). Either 80% or 20% of the descrambled sentences described hostile behaviors.	McCarthy et al.
Vignette (Srull & Wyer, 1979, Experiment 1)	Participants read a short story about a man named Ronald who behaved in a manner that could be seen as hostile (e.g., he told a beggar to find a job).	McCarthy et al.
Judgments of the vignette's protagonist (Srull & Wyer, 1979, Experiment 1)	Participants rated Ronald on 12 characteristics (e.g., unfriendly).	McCarthy et al.
Judgments of behaviors (Srull & Wyer, 1979, Experiment 1)	Participants judged the hostility of 15 behaviors (e.g., refusing to let a salesperson into one's house).	McCarthy et al.
Abstract reasoning (materials provided by C. Chabris)	Participants solved a 10-item nonverbal-intelligence task.	Filler
Priming (moral reminder)	Participants wrote as many of the Ten Commandments as they could remember or the names of 10 books they had read in high school.	Current
Matrix (cheating opportunity) (Mazar et al., 2008, Experiment 1)	Participants tried to find the numbers that added up exactly to 10 (e.g., 3.18 and 6.82) in as many of 20 matrices as time allowed. They then tore either a blank page or the matrix page out of the task booklet.	Current
Collection slip (Mazar et al., 2008, Experiment 1)	Participants reported how many matrices they had solved.	Current
Alternative Uses Test (Guilford, 1967)	Participants listed as many possible uses of a paper clip as they could think of.	Filler
Religiousness[a]	Participants used a scale from 1 (*not at all*) to 5 (*completely*) to answer three questions: "How religious are you?"; "To what extent do you believe in a God?"; and "To what extent do you believe in a punishing God?"	Current
Fatigue[a] (Profile of Mood States; McNair, Lorr, & Droppleman, 1971) and sleep	Participants rated their fatigue, by using a scale from 1 (*not at all*) to 5 (*extremely*) to indicate how much they felt worn out, fatigued, exhausted, sluggish, weary, and bushed; participants also reported how many hours they had slept the previous night.	Filler
Time estimation[a]	Participants estimated how much time they had taken in the timed tasks of this battery.	Current
HEXACO[a] (Ashton & Lee, 2009)	Participants completed this 60-item personality scale.	Filler

Note: This table lists the order of all of the tasks included in the combined procedure for the current RRR, on Mazar, Amir, and Ariely's (2008) Experiment 1, and for McCarthy et al.'s (2018, this issue) RRR, on Srull and Wyer's (1979) Experiment 1. All between-participants conditions were counterbalanced.

[a]These tasks were included to allow exploratory analyses of possible moderators of cheating. The religiousness task was included in the preregistered plan.

Two between-subjects variables were manipulated: priming task (recall the Ten Commandments vs. recall 10 books from high school) and presence or absence of the opportunity to cheat (i.e., whether or not the self-reported number of matrices solved could be verified). Participants who received the Ten Commandments prime read the following instructions: "For this next task, please write down as many of the 10 Commandments from the Bible as you remember. Please time yourself and spend no more than 2 minutes on this task." Participants who received the books prime read: "For this next task, please write down the names of 10 books that you read in high school. Please time yourself and spend no more than 2 minutes on this task." The combinations of the priming and cheating variables yielded four conditions that we refer to as the Commandments-cheat, books-cheat, Commandments-control, and books-control conditions.

The problem-solving task consisted of 20 matrices (half of which were unsolvable). Participants were told to allot 4 min to complete as many matrices as possible. The instructions on this page also noted that 2 participants, chosen at random from all participants in the study, would be paid $10 for each matrix they solved (or that they would be paid the equivalent in another currency, in the case of labs outside the United States). Participants in the cheat conditions were asked to tear out the page with the matrices and to keep it for themselves, handing in the remainder of the package that contained only the page on which they reported the number of matrices they had solved. Participants in the control conditions were asked to tear out a blank page (to mask the presence of other conditions in the testing session). Those participants submitted in their package both the page on which they reported the number of correctly solved matrices and the matrices sheet.

During protocol development and in consultation with Mazar et al., we identified several aspects of the original design that were not mentioned in the original article but that potentially were important. In all cases, we used the same procedures as in the original study. First, half of the 20 matrices were actually unsolvable, but participants were not told this. Second, the example matrix (Fig. 1) accompanying the written instructions showed two circled numbers adding to 10. However, the instructions did not specify that participants should circle no more than two numbers, and it was left to participants whether they tried to solve the matrices with sets of more than two numbers. Third, as the matrix task was self-timed, participants could cheat either by overreporting the number of correctly solved matrices or by taking more than the allotted 4 min and actually solving more matrices. In the latter case, participants would be cheating by violating the instructions rather than by inflating their performance report.

Differences from the original study

In the original study, participants in the cheat conditions, but not those in the control conditions, tore out a page from the booklet (i.e., the page with the matrices). If some participants in a session tore out pages and others did not, that difference could suggest the presence of multiple conditions to the participants, and it could reveal a given participant's condition to the experimenters. To avoid this possible unblinding, we had participants in the control conditions tear out a blank page from the booklet.

In the original study, the number of matrices solved was defined differently for the cheat and control conditions. In the cheat conditions, for which participants kept the matrix page, the dependent measure was the self-reported number of matrices solved. In the control conditions, the experimenters coded the submitted matrix page to determine whether or not participants correctly circled the two numbers adding to 10 in each matrix, and the dependent measure used for analysis was the total number of correctly solved matrices (N. Mazar, personal communication, April 5, 2018). The analysis for this RRR used the self-reported total number of solved matrices for both the cheat and the control conditions to ensure that differences between conditions could not be attributed to differences in the outcome measure. We also conducted exploratory analyses in which we used the number of correctly solved matrices (coded by the experimenters from the matrix pages) as the dependent variable for the control conditions.

In the original study, participants were not explicitly instructed to circle the numbers, but were told only to mark the "Got it" box below each matrix they solved. Mazar et al. noted that most participants in the control conditions followed the example set by the sample problem and spontaneously circled the two numbers adding to 10. To ensure our ability to verify the accuracy of the self-reports, in the control conditions we added explicit instructions to circle the numbers adding up to 10.[2]

We added "from the Bible" to the instructions in the Commandments conditions because pilot testing showed that some participants (e.g., nonreligious people) might not know what was meant by the "10 Commandments." The RRR protocol also added text to the example matrix problem to ensure that the task was clear to participants (i.e., "In the example to the right, 3.81 and 6.19 add up exactly to 10"). For the RRR tasks, we projected a stopwatch on a screen at the front of

the room rather than asking participants to use their own devices to time themselves. We did so both to standardize procedures and because not all testing rooms had clocks, smartphones might allow for cheating, and fewer students now carry watches than did when Mazar et al. conducted their experiment. Finally, the original authors have no record of the other tasks included in the testing battery. We selected a set of tasks, vetted by the original authors, that were not expected to influence performance on the primary task.

Inclusion criteria

To be included in the analyses, participants had to be part of a sample that was 20% to 80% female and had to be 18- to 25-year-old students at the time of testing. They also had to follow task instructions and to complete all the tasks necessary for this replication study. These last two criteria were underspecified in the preregistered protocol, and the lead labs and Editor clarified these criteria prior to examining the data and results. Not following task instructions included not having torn out the matrix or blank page or reporting having solved more than 20 matrices. Participants who listed no books or Commandments and also reported spending no time recalling them were excluded for not having completed all the tasks. Note that participants were not excluded for not having completed other tasks in the packet. Finally, data were excluded when the experimenter did not administer the tasks correctly or when a testing session included fewer than 50 participants (to ensure adequate anonymity, as in the original study).

Data-blind exceptions to the protocol were allowed (e.g., we allowed labs to recruit samples that had less than 20% males). All exceptions are listed in Table 2, which also reports the language used at each lab, the number of participants tested and the number included in analyses for each lab, and descriptive statistics for the dependent variable. Note that all data, both excluded and included, are available on the OSF project page (https://osf.io/vxz7q/). Labs indicated in their data file whether or not participants' data were included and, if not, the reason for the exclusion.

Results

The R scripts (R Version 3.4.3; R Core Team, 2013) for the data analysis were written during the data-collection phase and registered before viewing the data (see https://osf.io/vxz7q/). Prior to the primary data analysis, a data-integrity script checked for potential errors in data entry or coding, and individual labs were asked to clarify or resolve potential errors. As is standard for RRR projects, the primary data analysis consisted of a random-effects meta-analysis (Simons, Holcombe, & Spellman, 2014). For each lab, the effect size was calculated as the difference in the mean number of solved matrices between the books-cheat condition and the Commandments-cheat condition. To examine whether differences in the religiousness of participants across labs moderated the size of the effect observed, we conducted a preregistered exploratory metaregression using the random-effects model (Thompson & Higgins, 2002). The analysis was based on the average response across three single-item religiousness measures (see Table 1; separate analyses for each of the three religiousness measures can be found on the OSF project page, https://osf.io/vxz7q/).

Primary analyses

The primary analyses included data from 4,674 participants from 19 laboratories that met all inclusion criteria or that were granted a data-blind exception. The exceptions included 7 labs with samples that were more than 80% female and 1 lab that allowed people up to 27 years of age to participate (exception granted but not needed).

Our primary analysis concerned the meta-analytic difference between the Commandments-cheat and the books-cheat conditions. That is, we asked whether people given a moral prime report solving fewer matrices than do those given a neutral prime when they are given the opportunity to cheat. In the original study, participants in the Commandments-cheat condition reported solving 1.45 fewer matrices than did those in the books-cheat condition. In our replication project, participants reported solving 0.11 more matrices in the Commandments-cheat condition than in the books-cheat condition (95% confidence interval, CI = [−0.09, 0.31]; Fig. 2). This corresponds to a Cohen's d of −0.04 (95% CI = [−0.12, 0.04]; the negative sign reflects that the effect was numerically in the opposite direction of the effect in the original study). Seven out of the 19 labs showed an effect numerically in the same direction as in the original study, but none of the 95% CIs for these labs excluded zero.

There was no heterogeneity across labs, $\tau^2 = 0$, $Q(18) = 13.16$, $p = .78$ (Borenstein, Hedges, Higgins, & Rothstein, 2009), and 0% of the observed variance in the effect sizes was attributable to systematic differences between labs (I^2). Together, these indices suggested that further analyses of moderation by religiousness were not warranted. For completeness, Figure 3 plots the moderation of the Ten Commandments effect by religiousness. The meta-regression showed no significant effect for religiousness, with the point estimate of the slope being 0.17, 95% CI = [−0.09, 0.43], $p = .20$.

Table 2. Descriptive Statistics and General Information for the Participating Labs

Lab	Country	Language	Sample size		Mean number of matrices reported solved		Data-blind exceptions to the inclusion criteria
			Tested	Included in analyses	Commandments-cheat condition	Books-cheat condition	
Acar	United Kingdom	British English	237	163	3.97 (2.93)	3.75 (2.48)	None
Aczel	Hungary	Hungarian	245	215	3.18 (2.28)	3.44 (2.77)	Omitted added instruction to circle the two numbers
Baskin	United States	American English	207	173	3.31 (2.21)	3.05 (3.19)	None
Birt	Canada	American English	234	205	3.04 (2.77)	2.63 (2.31)	Lower male-to-female ratio
Blatz	Germany	German	320	196	4.31 (3.47)	3.24 (3.47)	Lower male-to-female ratio; omitted added instruction to circle the two numbers
Evans	United States	American English	332	231	3.08 (2.88)	2.23 (2.31)	None
Ferreira-Santos	Portugal	Portuguese	291	211	2.67 (2.56)	2.85 (2.41)	None
González-Iraizoz	United Kingdom	British English	235	214	3.67 (3.05)	3.41 (2.46)	Lower male-to-female ratio
Holzmeister	Austria	German	274	246	7.02 (4.33)	5.91 (3.63)	Omitted added instruction to circle the two numbers
klein Selle & Rozmann	Israel	Hebrew	337	283	3.31 (2.40)	3.58 (2.66)	None
Koppel	Sweden	Swedish	263	236	3.11 (2.13)	2.73 (2.31)	Omitted added instruction to circle the two numbers
Laine	France	French	313	224	2.19 (2.09)	2.56 (2.32)	Lower male-to-female ratio; omitted added instruction to circle the two numbers
Loschelder	Germany	German	248	212	3.98 (1.79)	4.09 (2.27)	Omitted added instruction to circle the two numbers
McCarthy	United States	American English	318	218	3.40 (3.76)	2.83 (3.46)	None
Meijer	Netherlands	English	377	336	3.02 (1.52)	3.18 (2.33)	None
Özdoğru	Turkey	Turkish and English	365	237	4.45 (2.88)	3.26 (3.06)	Lower male-to-female ratio; omitted added instruction to circle the two numbers
Pennington	United Kingdom	British English	255	197	2.85 (2.01)	2.10 (1.59)	None
Roets	Belgium	Dutch	253	192	3.76 (2.45)	3.40 (2.13)	Lower male-to-female ratio
Suchotzki	Germany	German	256	240	3.83 (2.25)	3.83 (2.89)	Lower male-to-female ratio; omitted added instruction to circle the two numbers
Sutan	France	French and English	304	300	4.68 (1.89)	4.66 (2.38)	None
Tran	Austria	German	277	191	3.83 (3.18)	3.73 (2.23)	Omitted added instruction to circle the two numbers
Vanpaemel	Belgium	Dutch	288	227	3.61 (1.65)	3.44 (2.22)	None
Verschuere	Netherlands	Dutch	302	265	3.73 (2.30)	3.55 (1.97)	None
Wick	United States	American English	367	334	3.19 (2.50)	3.28 (3.60)	None
Wiggins	United States	American English	259	240	2.34 (2.19)	2.15 (1.79)	None
Across all labs			7,158	5,786	3.54 (2.74)	3.36 (2.73)	

Note: Numbers in parentheses are standard deviations.

Lab	Commandments-Cheat Condition: Mean Number of Matrices	Commandments-Cheat Condition: *n*	Books-Cheat Condition: Mean Number of Matrices	Books-Cheat Condition: *n*	Effect Size	95% CI
Original Result	2.77	56	4.22	50	–1.45	[–2.61, –0.29]
Laine	2.19	53	2.56	61	–0.37	[–1.18, 0.44]
klein Selle & Rozmann	3.31	70	3.58	69	–0.27	[–1.11, 0.58]
Aczel	3.18	55	3.44	54	–0.26	[–1.22, 0.69]
Ferreira-Santos	2.67	51	2.85	54	–0.19	[–1.14, 0.77]
Meijer	3.02	84	3.18	84	–0.15	[–0.75, 0.44]
Loschelder	3.98	59	4.09	55	–0.11	[–0.86, 0.65]
Wick	3.19	81	3.28	79	–0.09	[–1.06, 0.87]
Suchotzki	3.83	60	3.83	60	0.00	[–0.93, 0.93]
Sutan	4.68	34	4.66	87	0.02	[–0.79, 0.83]
Vanpaemel	3.61	56	3.44	59	0.17	[–0.55, 0.88]
Verschuere	3.73	64	3.55	69	0.18	[–0.55, 0.91]
Wiggins	2.34	62	2.15	61	0.19	[–0.51, 0.90]
González-Iraizoz	3.67	55	3.41	58	0.26	[–0.77, 1.28]
Koppel	3.11	62	2.73	62	0.39	[–0.40, 1.17]
Birt	3.04	53	2.63	49	0.41	[–0.58, 1.39]
McCarthy	3.40	50	2.83	46	0.57	[–0.87, 2.02]
Evans	3.08	53	2.23	57	0.85	[–0.13, 1.83]
Holzmeister	7.02	60	5.91	64	1.11	[–0.30, 2.52]
Özdoğru	4.45	38	3.26	66	1.19	[0.01, 2.37]
Meta-Analytic Average	3.55	1,100	3.35	1,194	0.11	[–0.09, 0.31]

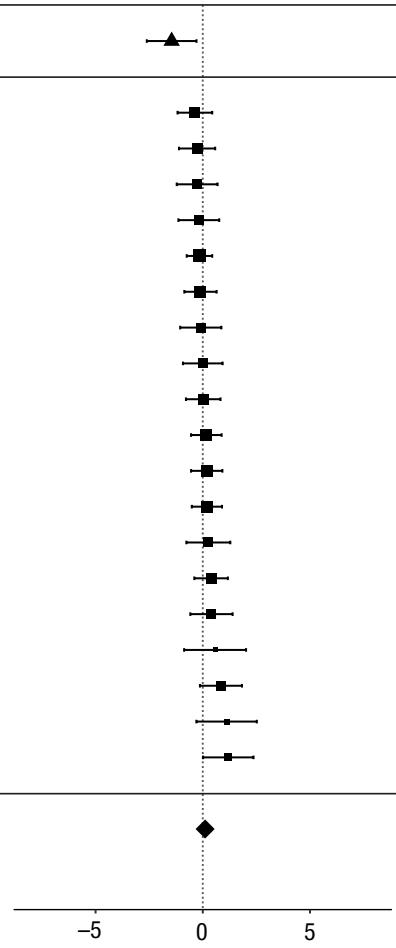

Fig. 2. Results of the primary analyses: forest plot of the difference between the Commandments-cheat and the books-cheat conditions in the self-reported number of matrices solved. For each of the 19 labs that met all the inclusion criteria, the figure shows the mean self-report and sample size in each condition. The labs are listed in order of the size of the difference between the conditions (Commandments-cheat condition minus books-cheat condition). The squares show the observed effect sizes, the error bars represent 95% confidence intervals (CIs), and the size of each square represents the magnitude of the standard error for the lab's effect (larger squares indicate less variability in the estimate). To the right, the figure shows the numerical values for the effect sizes and 95% CIs. At the top of the figure, the effect from Mazar, Amir, and Ariely's (2008) Experiment 1 is shown. The bottom row in the figure presents the unweighted means of the individual sample means and the outcome of a random-effects meta-analysis. Note that the meta-analytic estimate of the difference between conditions does not necessarily equal the difference between the means.

Ancillary analyses: other comparisons of interest

Mazar et al. (2008) predicted that a moral reminder would reduce cheating, and they found that it completely eliminated cheating. Consequently, in our replication project, we expected that the reported number of matrices solved would be comparable in the Commandments-cheat condition and the Commandments-control condition. In the original study, the difference between these conditions in number of matrices solved (Commandments-cheat condition minus Commandments-control condition) was −0.35 matrices. In our RRR project, the meta-analytic effect was 0.24 matrices (95% CI = [0.03, 0.44]), and there was no significant heterogeneity across labs, $\tau^2 = 0.01$, $I^2 = 4.48$, $Q(18) = 19.23$, $p = .38$ (Fig. 4).

We also predicted that the Commandments prime would not have an effect among participants without an opportunity to cheat. That is, we expected the reported number of matrices solved to be comparable in the Commandments-control condition and the books-control condition. In the original study, the difference between these conditions (Commandments-control condition minus books-control condition) was 0.05

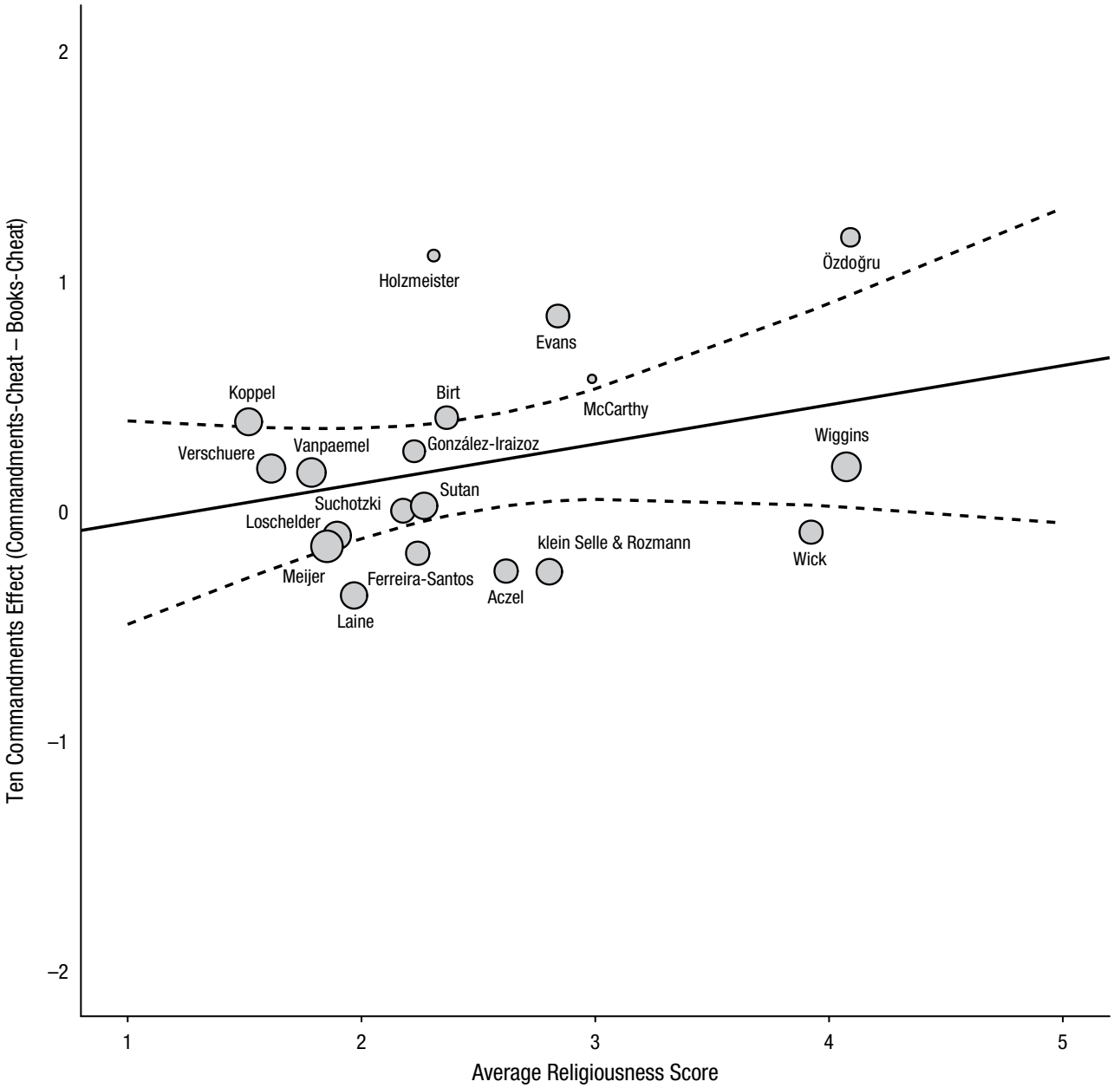

Fig. 3. Moderation of the Ten Commandments effect by religiousness for the 19 labs included in the primary analyses. The scatterplot shows the magnitude of the effect and the average religiousness score for each lab. The solid line is the best-fitting regression line, and the dashed lines mark the 95% confidence band around the regression line. The size of each circle represents the magnitude of the standard error for the lab's effect (larger circles indicate less variability).

matrices. In the current replication project, the meta-analytic effect was 0.01 matrices (95% CI = [–0.19, 0.20]), and there was no heterogeneity across labs, $\tau^2 = 0$, $I^2 = 0$, $Q(18) = 15.30$, $p = 0.64$ (Fig. 5).

Our third prediction was that the books prime would not reduce the tendency to cheat. That is, we expected the reported number of matrices solved to be higher in the books-cheat condition than in the books-control condition. In the original study, this difference (books-cheat condition minus books-control condition) was 1.16 matrices. In the replication project, the meta-analytic effect was 0.15 matrices (95% CI = [–0.03, 0.34]), and there was no heterogeneity across labs, $\tau^2 = 0$, $I^2 = 0$, $Q(18) = 14.00$, $p = .73$ (Fig. 6).

Finally, we predicted that the difference between the cheat and control conditions would be greater in the books conditions than in the Commandments conditions. In the original study, the cheat-control difference was 1.16 matrices in the books conditions and –0.35 matrices in the Commandments conditions (difference = 1.51). In

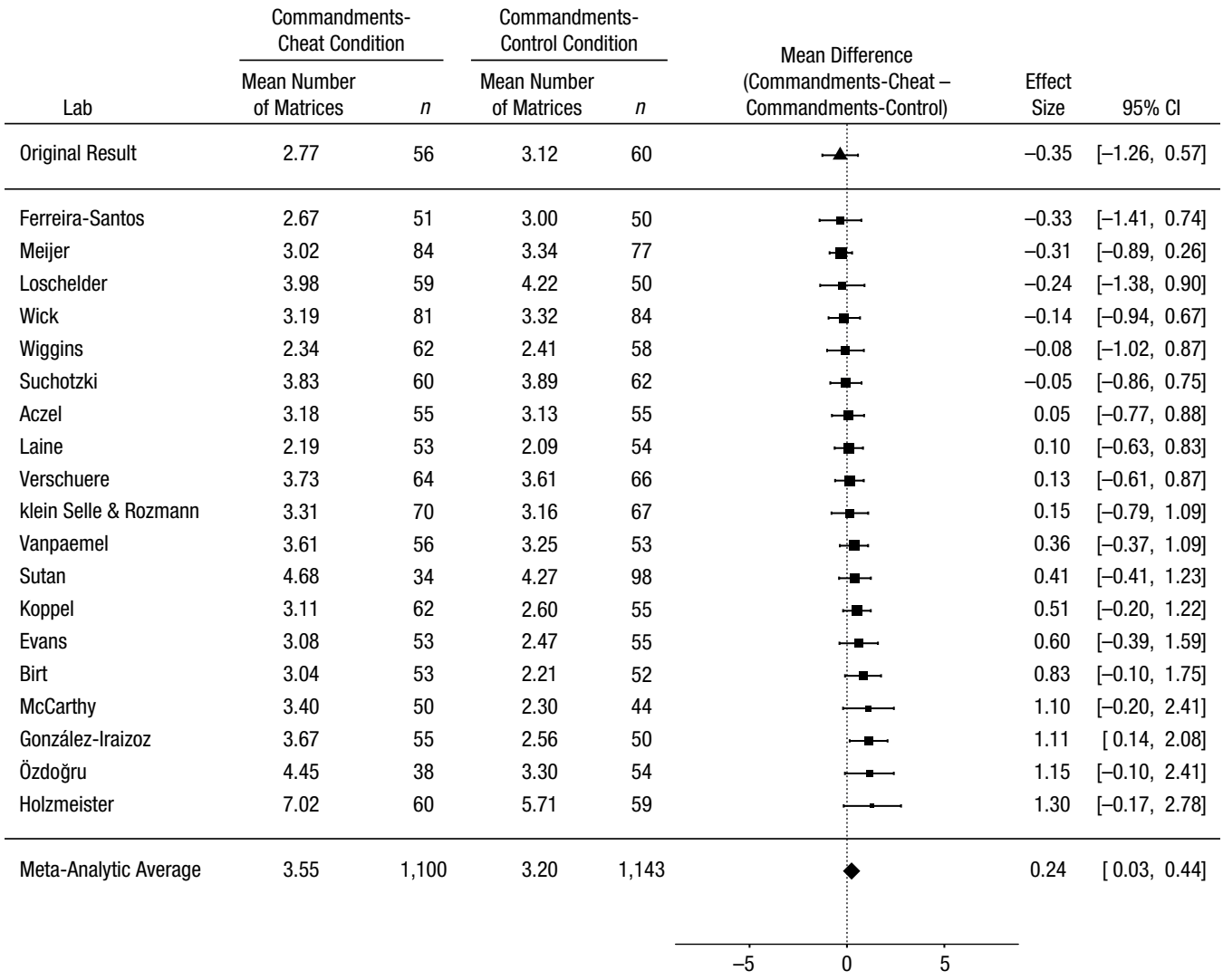

Fig. 4. Results of the ancillary analyses including the 19 labs that met all the inclusion criteria: forest plot of the difference between the Commandments-cheat and the Commandments-control conditions in the self-reported number of matrices solved. For each lab, the figure shows the mean self-report and sample size in each condition. The labs are listed in order of the size of the difference between the conditions (Commandments-cheat condition minus Commandments-control condition). The squares show the observed effect sizes, the error bars represent 95% confidence intervals (CIs), and the size of each square represents the magnitude of the standard error for the lab's effect (larger squares indicate less variability in the estimate). To the right, the figure shows the numerical values for the effect sizes and 95% CIs. At the top of the figure, the effect from Mazar, Amir, and Ariely's (2008) Experiment 1 is shown. The bottom row in the figure presents the unweighted means of the individual sample means and the outcome of a random-effects meta-analysis. Note that the meta-analytic estimate of the difference between conditions does not necessarily equal the difference between the means.

our replication project, the cheat-control difference was 0.11 matrices in the books conditions and 0.35 matrices in the Commandments conditions (difference = −0.11, 95% CI = [−0.39, 0.17]), and there was no heterogeneity across labs, $\tau^2 = 0$, $I^2 = 0$, $Q(18) = 16.21$, $p = .58$ (Fig. 7).

Ancillary analyses: primary outcome measure across all labs

We repeated the main analysis including the data of all 25 laboratories (total $N = 5{,}786$) that submitted data for this RRR study. Participants reported solving 0.17 more matrices in the Commandments-cheat condition than in the books-cheat condition (95% CI = [−0.00, 0.35]; Fig. 8). Seven out of the 25 labs found an effect in the same direction as in the original study, but the 95% CI did not exclude zero in any of these cases. Cochran's Q revealed no heterogeneity across the 25 labs, $\tau^2 = 0$, $Q(24) = 17.46$, $p = .83$, and I^2 indicated that about 0% of the observed variance in effect sizes was caused by systematic differences between labs. A metaregression showed no significant effect for religiousness; the point estimate of the slope was 0.12, 95% CI = [−0.13, 0.37], $p = .36$ (see Fig. 9).

Lab	Commandments-Control Condition: Mean Number of Matrices	Commandments-Control Condition: *n*	Books-Control Condition: Mean Number of Matrices	Books-Control Condition: *n*	Effect Size	95% CI
Original Result	3.12	60	3.06	63	0.05	[–0.78, 0.89]
Ferreira-Santos	3.00	50	2.14	56	0.86	[–0.10, 1.81]
Vanpaemel	3.25	53	2.73	59	0.52	[–0.27, 1.31]
Wick	3.32	84	2.81	90	0.51	[–0.32, 1.34]
Wiggins	2.41	58	1.93	59	0.48	[–0.36, 1.33]
Verschuere	3.61	66	3.29	66	0.32	[–0.35, 0.98]
Evans	2.47	55	2.26	66	0.22	[–0.54, 0.97]
Holzmeister	5.71	59	5.59	63	0.12	[–1.16, 1.41]
Suchotzki	3.89	62	3.90	58	–0.01	[–0.88, 0.86]
González-Iraizoz	2.56	50	2.63	51	–0.07	[–0.84, 0.70]
klein Selle & Rozmann	3.16	67	3.25	77	–0.08	[–0.94, 0.77]
Meijer	3.34	77	3.47	91	–0.13	[–0.78, 0.51]
Aczel	3.13	55	3.31	51	–0.19	[–0.97, 0.60]
Laine	2.09	54	2.34	56	–0.25	[–0.92, 0.42]
Sutan	4.27	98	4.54	81	–0.28	[–0.96, 0.41]
Koppel	2.60	55	2.95	57	–0.35	[–1.21, 0.51]
Loschelder	4.22	50	4.63	48	–0.41	[–1.79, 0.98]
Özdoğru	3.30	54	3.77	79	–0.48	[–1.56, 0.61]
McCarthy	2.30	44	2.99	78	–0.69	[–1.82, 0.44]
Birt	2.21	52	3.08	51	–0.87	[–2.10, 0.37]
Meta-Analytic Average	3.20	1,143	3.24	1,237	0.01	[–0.19, 0.20]

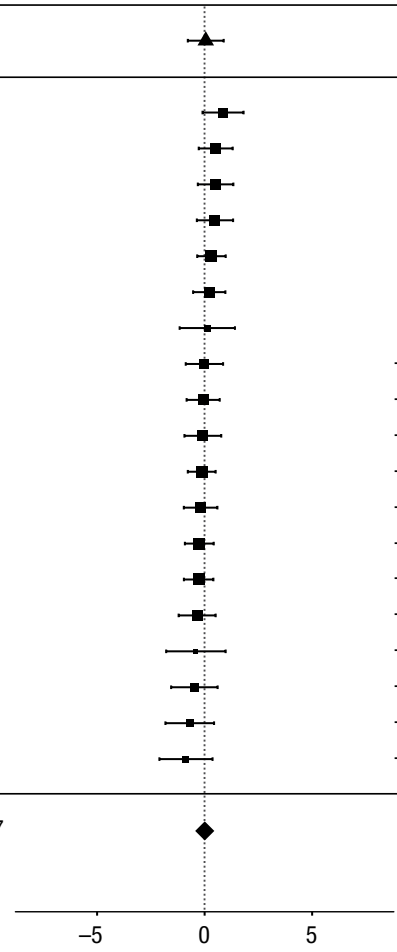

Fig. 5. Results of the ancillary analyses including the 19 labs that met all the inclusion criteria: forest plot of the difference between the Commandments-control and the books-control conditions in the self-reported number of matrices solved. For each lab, the figure shows the mean self-report and sample size in each condition. The labs are listed in order of the size of the difference between the conditions (Commandments-control condition minus books-control condition). The squares show the observed effect sizes, the error bars represent 95% confidence intervals (CIs), and the size of each square represents the magnitude of the standard error for the lab's effect (larger squares indicate less variability in the estimate). To the right, the figure shows the numerical values for the effect sizes and 95% CIs. At the top of the figure, the effect from Mazar, Amir, and Ariely's (2008) Experiment 1 is shown. The bottom row in the figure presents the unweighted means of the individual sample means and the outcome of a random-effects meta-analysis. Note that the meta-analytic estimate of the difference between conditions does not necessarily equal the difference between the means.

Ancillary analyses: primary outcome measure across labs that strictly met all inclusion criteria

Finally, we repeated the main analysis including only the data of the 10 laboratories (total n = 2,645) that strictly met all a priori inclusion criteria (no exceptions allowed). Participants reported solving 0.07 more matrices in the Commandments-cheat condition than in the books-cheat condition (95% CI = [–0.18, 0.33]; see Fig. S1 in the Supplemental Material or on the OSF project page, at https://osf.io/vxz7q/). Four out of the 10 labs found an effect in the same direction as in the original study, but none of these effects had a 95% CI that excluded zero. Cochran's Q revealed no heterogeneity across labs, $\tau^2 = 0$, $Q(9) = 4.71$, $p = .86$, and I^2 indicated that 0% of the observed variance in effect sizes was caused by systematic differences between labs. A metaregression showed no significant effect for religiousness; the point estimate of the slope was 0.05,

Lab	Books-Cheat Condition: Mean Number of Matrices	Books-Cheat Condition: *n*	Books-Control Condition: Mean Number of Matrices	Books-Control Condition: *n*	Effect Size	95% CI
Original Result	4.22	50	3.06	63	1.16	[0.06, 2.25]
González-Iraizoz	3.41	58	2.63	51	0.79	[–0.05, 1.62]
Vanpaemel	3.44	59	2.73	59	0.71	[–0.07, 1.49]
Ferreira-Santos	2.85	54	2.14	56	0.71	[–0.11, 1.52]
Wick	3.28	79	2.81	90	0.47	[–0.51, 1.45]
klein Selle & Rozmann	3.58	69	3.25	77	0.33	[–0.41, 1.08]
Holzmeister	5.91	64	5.59	63	0.32	[–0.89, 1.53]
Verschuere	3.55	69	3.29	66	0.26	[–0.39, 0.92]
Laine	2.56	61	2.34	56	0.22	[–0.54, 0.98]
Wiggins	2.15	61	1.93	59	0.22	[–0.35, 0.78]
Aczel	3.44	54	3.31	51	0.13	[–0.78, 1.04]
Sutan	4.66	87	4.54	81	0.11	[–0.56, 0.79]
Evans	2.23	57	2.26	66	–0.03	[–0.77, 0.71]
Suchotzki	3.83	60	3.90	58	–0.06	[–1.05, 0.92]
McCarthy	2.83	46	2.99	78	–0.16	[–1.45, 1.13]
Koppel	2.73	62	2.95	57	–0.22	[–1.14, 0.70]
Meijer	3.18	84	3.47	91	–0.29	[–0.96, 0.37]
Birt	2.63	49	3.08	51	–0.45	[–1.73, 0.84]
Özdoğru	3.26	66	3.77	79	–0.51	[–1.50, 0.48]
Loschelder	4.09	55	4.63	48	–0.53	[–1.62, 0.56]
Meta-Analytic Average	3.35	1,194	3.24	1,237	0.15	[–0.03, 0.34]

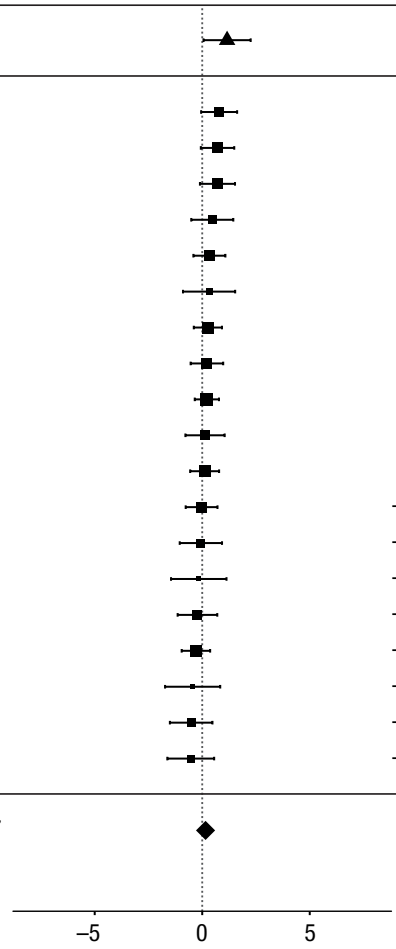

Fig. 6. Results of the ancillary analyses including the 19 labs that met all the inclusion criteria: forest plot of the difference between the books-cheat and the books-control conditions in the self-reported number of matrices solved. For each lab, the figure shows the mean self-report and sample size in each condition. The labs are listed in order of the size of the difference between the conditions (books-cheat condition minus books-control condition). The squares show the observed effect sizes, the error bars represent 95% confidence intervals (CIs), and the size of each square represents the magnitude of the standard error for the lab's effect (larger squares indicate less variability in the estimate). To the right, the figure shows the numerical values for the effect sizes and 95% CIs. At the top of the figure, the effect from Mazar, Amir, and Ariely's (2008) Experiment 1 is shown. The bottom row in the figure presents the unweighted means of the individual sample means and the outcome of a random-effects meta-analysis. Note that the meta-analytic estimate of the difference between conditions does not necessarily equal the difference between the means.

95% CI = [–0.23, 0.33], $p = .72$ (see Fig. S2 in the Supplemental Material or at https://osf.io/vxz7q/).

Exploratory analyses

To maximize power, we conducted all exploratory analyses on data from all 25 labs. The original study found a higher number of matrices solved in the books-cheat condition than in the books-control condition, an effect that was attributed to cheating in the absence of a moral reminder when participants could cheat without risk of being caught (i.e., participants in the books-cheat condition ripped out the matrix page from their packet). We did not find this difference in the primary analysis, but that analysis and the original study used different dependent measures for the control condition: Our primary analysis used the self-reported total number of solved matrices, and the original study used the actual number as verified by the experimenter. When we analyzed the data in the same way as in the original study, comparing self-reports for the cheat condition with the verified number of correctly solved

Lab	Books Conditions: Cheat-Control Effect	Books Conditions: *n*	Commandments Conditions: Cheat-Control Effect	Commandments Conditions: *n*	Effect Size	95% CI
Original Result	1.16	113	–0.35	116	1.51	[0.08, 2.93]
Ferreira-Santos	0.71	110	–0.33	101	1.04	[–0.31, 2.39]
Wick	0.47	169	–0.14	165	0.60	[–0.67, 1.87]
Vanpaemel	0.71	118	0.36	109	0.35	[–0.71, 1.41]
Wiggins	0.22	120	–0.08	120	0.29	[–0.81, 1.39]
klein Selle & Rozmann	0.33	146	0.15	137	0.18	[–1.02, 1.39]
Verschuere	0.26	135	0.13	130	0.13	[–0.85, 1.12]
Laine	0.22	117	0.10	107	0.12	[–0.93, 1.17]
Aczel	0.13	105	0.05	110	0.08	[–1.16, 1.31]
Meijer	–0.29	175	–0.31	161	0.02	[–0.86, 0.90]
Suchotzki	–0.06	118	–0.05	122	–0.01	[–1.28, 1.26]
Loschelder	–0.53	103	–0.24	109	–0.30	[–1.87, 1.28]
Sutan	0.11	168	0.41	132	–0.30	[–1.36, 0.76]
González-Iraizoz	0.79	109	1.11	105	–0.33	[–1.61, 0.96]
Evans	–0.03	123	0.60	108	–0.63	[–1.87, 0.60]
Koppel	–0.22	119	0.51	117	–0.73	[–1.90, 0.43]
Holzmeister	0.32	127	1.30	119	–0.99	[–2.89, 0.92]
McCarthy	–0.16	124	1.10	94	–1.27	[–3.10, 0.57]
Birt	–0.45	100	0.83	105	–1.27	[–2.85, 0.31]
Özdoğru	–0.51	145	1.15	92	–1.67	[–3.26, –0.07]
Meta-Analytic Average	0.11	2,431	0.35	2,243	–0.11	[–0.39, 0.17]

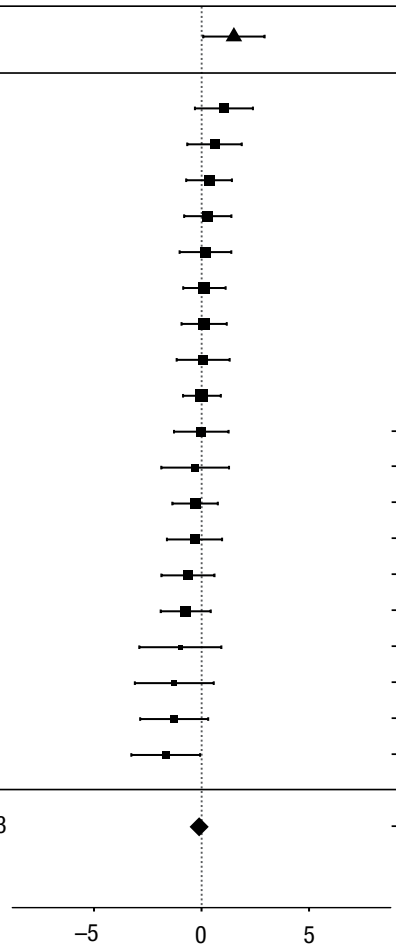

Fig. 7. Results of the ancillary analyses including the 19 labs that met all the inclusion criteria: forest plot of the difference between the cheat-control effect in the books conditions (books-cheat minus books-control) and the cheat-control effect in the Commandments conditions (Commandments-cheat minus Commandments-control). For each lab, the figure shows the mean effect and sample size in each condition. The labs are listed in order of the size of the effect. The squares show the observed effect sizes, the error bars represent 95% confidence intervals (CIs), and the size of each square represents the magnitude of the standard error for the lab's effect (larger squares indicate less variability in the estimate). To the right, the figure shows the numerical values for the effect sizes and 95% CIs. At the top of the figure, the effect from Mazar, Amir, and Ariely's (2008) Experiment 1 is shown. The bottom row in the figure presents the unweighted means of the individual sample means and the outcome of a random-effects meta-analysis. Note that the meta-analytic estimate of the difference between conditions does not necessarily equal the difference between the means.

matrices for the control condition, we found a similar (but smaller) effect: The reported number of matrices solved in the books-cheat condition was 0.56 higher than the actual number of matrices solved in the books-control condition (95% CI = [0.35, 0.77]; see Fig. 10). Priming with the Ten Commandments did not result in reduced cheating when we used this dependent measure for the control condition, though: The reported number of matrices solved in the Commandments-cheat condition was 0.83 higher than the actual number of matrices solved in the Commandments-control condition (95% CI = [0.59, 1.06]; see Fig. S3 in the Supplemental Material or at https://osf.io/vxz7q/).

Discussion

Mazar et al. (2008, Experiment 1) reported that recalling the Ten Commandments—a moral reminder—reduced cheating more than did recalling 10 books from high school. This project replicated their procedures but did not find evidence of reduced cheating following the moral reminder. The results from the primary analysis

Lab	Commandments-Cheat Condition: Mean Number of Matrices	Commandments-Cheat Condition: *n*	Books-Cheat Condition: Mean Number of Matrices	Books-Cheat Condition: *n*	Effect Size	95% CI
Original Result	2.77	56	4.22	50	–1.45	[–2.61, –0.29]
Laine	2.19	53	2.56	61	–0.37	[–1.18, 0.44]
klein Selle & Rozmann	3.31	70	3.58	69	–0.27	[–1.11, 0.58]
Aczel	3.18	55	3.44	54	–0.26	[–1.22, 0.69]
Ferreira-Santos	2.67	51	2.85	54	–0.19	[–1.14, 0.77]
Meijer	3.02	84	3.18	84	–0.15	[–0.75, 0.44]
Loschelder	3.98	59	4.09	55	–0.11	[–0.86, 0.65]
Wick	3.19	81	3.28	79	–0.09	[–1.06, 0.87]
Suchotzki	3.83	60	3.83	60	0.00	[–0.93, 0.93]
Sutan	4.68	34	4.66	87	0.02	[–0.79, 0.83]
Tran	3.83	53	3.73	45	0.10	[–0.98, 1.17]
Vanpaemel	3.61	56	3.44	59	0.17	[–0.55, 0.88]
Verschuere	3.73	64	3.55	69	0.18	[–0.55, 0.91]
Wiggins	2.34	62	2.15	61	0.19	[–0.51, 0.90]
Acar	3.97	34	3.75	44	0.22	[–1.01, 1.45]
Baskin	3.31	36	3.05	41	0.26	[–0.96, 1.47]
González-Iraizoz	3.67	55	3.41	58	0.26	[–0.77, 1.28]
Roets	3.76	45	3.40	53	0.36	[–0.56, 1.28]
Koppel	3.11	62	2.73	62	0.39	[–0.40, 1.17]
Birt	3.04	53	2.63	49	0.41	[–0.58, 1.39]
McCarthy	3.40	50	2.83	46	0.57	[–0.87, 2.02]
Pennington	2.85	41	2.10	52	0.76	[0.01, 1.51]
Evans	3.08	53	2.23	57	0.85	[–0.13, 1.83]
Blatz	4.31	42	3.24	41	1.07	[–0.43, 2.56]
Holzmeister	7.02	60	5.91	64	1.11	[–0.30, 2.52]
Özdoğru	4.45	38	3.26	66	1.19	[0.01, 2.37]
Meta-Analytic Average	3.58	1,351	3.31	1,470	0.17	[–0.00, 0.35]

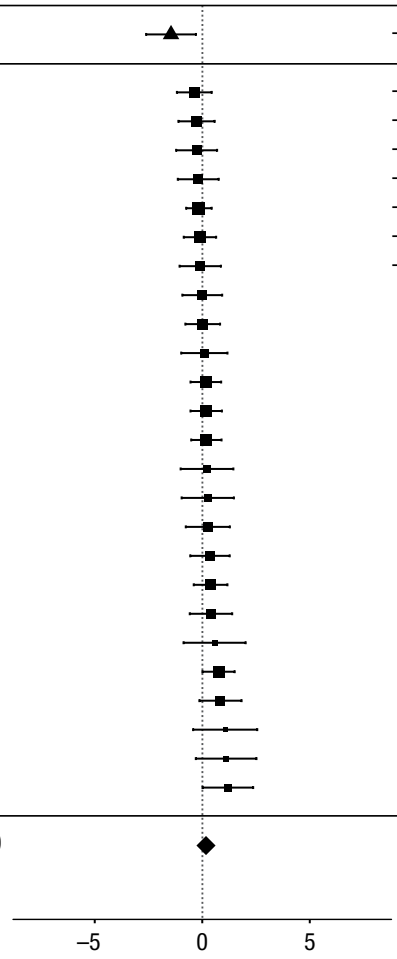

Fig. 8. Results of the ancillary analyses including all 25 labs that contributed to the replication project: forest plot of the difference between the Commandments-cheat and the books-cheat conditions in the self-reported number of matrices solved. For each lab, the figure shows the mean self-report and sample size in each condition. The labs are listed in order of the size of the difference between the conditions (Commandments-cheat condition minus books-cheat condition). The squares show the observed effect sizes, the error bars represent 95% confidence intervals (CIs), and the size of each square represents the magnitude of the standard error for the lab's effect (larger squares indicate less variability in the estimate). To the right, the figure shows the numerical values for the effect sizes and 95% CIs. At the top of the figure, the effect from Mazar, Amir, and Ariely's (2008) Experiment 1 is shown. The bottom row in the figure presents the unweighted means of the individual sample means and the outcome of a random-effects meta-analysis. Note that the meta-analytic estimate of the difference between conditions does not necessarily equal the difference between the means.

(19 labs, $n = 4{,}674$) and two ancillary analyses (lenient inclusion criterion: 25 labs, $N = 5{,}786$; strict inclusion criterion: 10 labs, $n = 2{,}645$) were consistent in showing a Ten Commandments effect close to zero. The effect was comparably small across labs, with no heterogeneity, which suggests that the differences among labs are consistent with sampling error rather than unexplained moderation.[3] For 24 laboratories, the confidence interval for this primary effect included zero, and the remaining lab found an effect in the opposite direction.

Given the discrepancy in findings, the differences between the replication project and the original study require consideration. A first difference is that the original study was run more than 10 years ago, at an elite university. The perceived rewards, perceived probability of getting caught cheating, and perceived consequences of getting caught may have been different for the participants in the current project. A second difference is the composition of the task battery that preceded the tasks for this study. Given that no record of the tasks in the original battery was kept, we selected

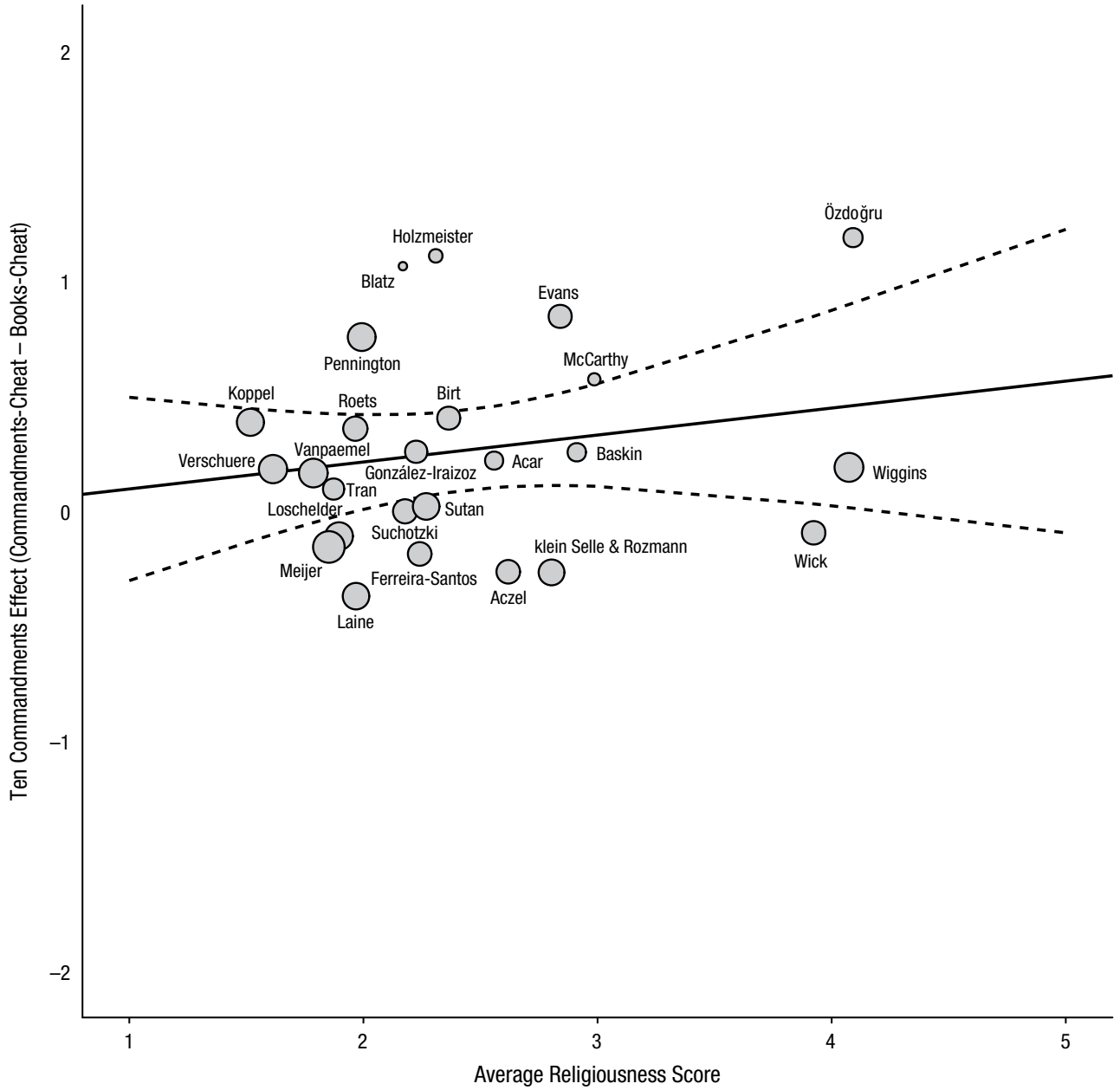

Fig. 9. Moderation of the Ten Commandments effect by religiousness for all 25 labs that contributed to the replication project. The scatterplot shows the magnitude of the effect and the average religiousness score for each lab. The solid line is the best-fitting regression line, and the dashed lines mark the 95% confidence band around the regression line. The size of each circle represents the magnitude of the standard error for the lab's effect (larger circles indicate less variability).

new tasks, and it is possible that the different tasks used in the two studies affected the extent of cheating observed. However, the tasks we selected were unrelated to the manipulation, and the lead authors and original authors agreed that there was no a priori reason to predict that the chosen tasks would interfere with the manipulation or outcome measure.

Mazar et al. (2008) reported a difference of 1.16 matrices solved between the books-cheat condition and the books-control condition, and they attributed this difference to cheating when participants were given the opportunity to do so with impunity and in the absence of a moral reminder. We found no difference between these conditions when using the self-reported number of matrices solved as the outcome measure, but found a similar difference in an exploratory analysis that compared self-reports in the cheat condition and verified correct responses in the control condition (as in the original study). However, this difference might have resulted from differences in those measures rather than

Lab	Books-Cheat Condition: Mean Number of Matrices	Books-Cheat Condition: *n*	Books-Control Condition: Mean Number of Matrices	Books-Control Condition: *n*	Effect Size	95% CI
Original Result	4.22	50	3.06	63	1.16	[0.06, 2.25]
Holzmeister	5.91	64	4.33	63	1.57	[0.42, 2.72]
Aczel	3.44	54	2.14	51	1.31	[0.40, 2.21]
Wick	3.28	79	2.04	90	1.23	[0.35, 2.12]
Laine	2.56	61	1.35	55	1.21	[0.48, 1.94]
Baskin	3.05	41	2.02	51	1.03	[–0.08, 2.14]
González-Iraizoz	3.41	58	2.39	49	1.03	[0.22, 1.83]
Vanpaemel	3.44	59	2.58	59	0.86	[0.08, 1.65]
Ferreira-Santos	2.85	54	2.00	56	0.85	[0.03, 1.67]
Birt	2.63	49	1.82	51	0.81	[0.01, 1.61]
klein Selle & Rozmann	3.58	69	2.79	77	0.79	[0.04, 1.54]
McCarthy	2.83	46	2.05	78	0.77	[–0.33, 1.88]
Wiggins	2.15	61	1.41	58	0.73	[0.18, 1.28]
Sutan	4.66	87	3.99	80	0.67	[0.03, 1.31]
Evans	2.23	57	1.67	66	0.56	[–0.15, 1.28]
Özdoğru	3.26	66	2.76	79	0.50	[–0.42, 1.42]
Verschuere	3.55	69	3.15	66	0.40	[–0.26, 1.06]
Roets	3.40	53	3.11	47	0.29	[–0.56, 1.14]
Koppel	2.73	62	2.46	57	0.27	[–0.45, 0.99]
Tran	3.73	45	3.51	41	0.22	[–0.75, 1.19]
Blatz	3.24	41	3.04	55	0.21	[–0.96, 1.37]
Suchotzki	3.83	60	3.73	55	0.11	[–0.83, 1.04]
Acar	3.75	44	3.64	45	0.11	[–0.91, 1.12]
Meijer	3.18	84	3.12	91	0.06	[–0.59, 0.70]
Loschelder	4.09	55	4.34	47	–0.25	[–1.16, 0.67]
Pennington	2.10	52	2.91	54	–0.81	[–1.52, –0.10]
Meta-Analytic Average	3.31	1,470	2.73	1,521	0.56	[0.35, 0.77]

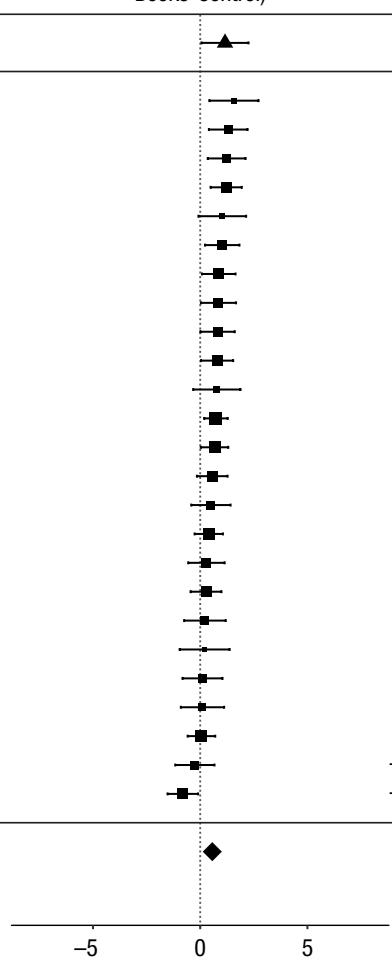

Fig. 10. Results of the exploratory analyses including all 25 labs that contributed to the replication project: forest plot of the difference between the number of matrices reported solved in the books-cheat condition and the number of matrices actually solved in the books-control condition. For each lab, the figure shows the mean number of matrices and sample size in each condition. The labs are listed in order of the size of the difference between the conditions (books-cheat condition minus books-control condition). The squares show the observed effect sizes, the error bars represent 95% confidence intervals (CIs), and the size of each square represents the magnitude of the standard error for the lab's effect (larger squares indicate less variability in the estimate). To the right, the figure shows the numerical values for the effect sizes and 95% CIs. At the top of the figure, the effect from Mazar, Amir, and Ariely's (2008) Experiment 1 is shown. The bottom row in the figure presents the unweighted means of the individual sample means and the outcome of a random-effects meta-analysis. Note that the meta-analytic estimate of the difference between conditions does not necessarily equal the difference between the means.

differences in cheating. For instance, if participants did not reliably circle the two numbers for each matrix, their number of verified correct responses would have been lower than their self-reported total, which would have resulted in a difference between the books-cheat condition and the books-control condition. Moreover, the difference observed using that dependent measure was greater rather than smaller (as it had been in the original study) when participants were primed with the Ten Commandments recall task, so our results are inconsistent with reduced cheating following a moral prime.

Future studies of the impact of moral reminders would benefit from using tasks that provide unambiguous evidence of cheating. Examples of such tasks include Gneezy's (2005) deception game, in which participants can maximize self-profit by duping another player, and variants of the coin-toss task that track participants' prediction before they can maximize profit by falsely claiming to have correctly predicted the result of the task (Peer, Acquisti, & Shalvi, 2014). Given skepticism about the effectiveness of religious priming (van Elk et al., 2015), future tests of the self-concept maintenance theory might

further benefit from exploring the effectiveness of non-religious moral primes (e.g., an honor pledge; Mazar et al., 2008) to evaluate whether they have a stronger influence on the proposed balance between maximizing self-profit and feeling moral.

In sum, we did not observe the predicted reduction in cheating following priming with the Ten Commandments. These results call into question the effectiveness of using the Ten Commandments as a moral prime to reduce cheating.

Appendix: Author Affiliations

(The Supplemental Material includes an additional appendix with a one-paragraph summary for each lab that specifies any departures from the protocol or from their own preregistered plan, as well as which analyses included the data from that lab. This Lab Implementation Appendix is also available at https://osf.io/vxz7q/).

Lead Labs

Randy J. McCarthy, Northern Illinois University
John J. Skowronski, Northern Illinois University

Bruno Verschuere, University of Amsterdam
Ariane Jim, University of Amsterdam, now at Ghent University

Ewout H. Meijer, Maastricht University
Katherine Hoogesteyn, Maastricht University
Robin Orthey, Maastricht University and University of Portsmouth

Contributing Labs
(Alphabetical by last name of first author)

Oguz A. Acar, City, University of London
Irene Scopelliti, City, University of London

Balazs Aczel, Institute of Psychology, Elte Eötvös Loránd University
Bence E. Bakos, Institute of Psychology, Elte Eötvös Loránd University
Marton Kovacs, Institute of Psychology, Elte Eötvös Loránd University
Peter Szecsi, Institute of Psychology, Elte Eötvös Loránd University

Ernest Baskin, Haub School of Business, Saint Joseph's University
Sean P. Coary, Haub School of Business, Saint Joseph's University

Angie R. Birt, Mount Saint Vincent University

Lisa Blatz, University of Cologne
Jan Crusius, University of Cologne

Jacqueline R. Evans, Florida International University
Keith Wylie, Florida International University
Steve D. Charman, Florida International University

Fernando Ferreira-Santos, University of Porto
Fernando Barbosa, University of Porto
Rita Pasion, University of Porto

Marta González-Iraizoz, University of Warwick
Andrea Isoni, University of Warwick
Elliot A. Ludvig, University of Warwick

Felix Holzmeister, University of Innsbruck
Juergen Huber, University of Innsbruck
Michael Kirchler, University of Innsbruck

Nathalie klein Selle, Hebrew University of Jerusalem
Noa Feldman, Hebrew University of Jerusalem
Gershon Ben-Shakhar, Hebrew University of Jerusalem
Nir Rozmann, Bar-Ilan University
Galit Nahari, Bar-Ilan University

Lina Koppel, Linköping University
Gustav Tinghög, Linköping University
Daniel Västfjäll, Linköping University and Decision Research, Eugene, Oregon

Tei Laine, Université Grenoble Alpes
Kévin Vezirian, Université Grenoble Alpes
Laurent Bègue, Université Grenoble Alpes

David D. Loschelder, Leuphana University of Lueneburg
Mario Mechtel, Leuphana University of Lueneburg

Asil Ali Özdoğru, Üsküdar University
Ezgi Yıldız, Üsküdar University

Charlotte R. Pennington, University of the West of England
Neil M. McLatchie, Lancaster University
Lara Warmelink, Lancaster University

Arne Roets, Ghent University
Alain Van Hiel, Ghent University

Kristina Suchotzki, University of Würzburg
Matthias Gamer, University of Würzburg

Angela Sutan, Université Bourgogne Franche-Comté, Burgundy School of Business - CEREN
Frank Lentz, Université Bourgogne Franche-Comté, Burgundy School of Business - CEREN
Jean-Christian Tisserand, Université Bourgogne Franche-Comté, Burgundy School of Business - CEREN
Eli Spiegelman, Université Bourgogne Franche-Comté, Burgundy School of Business - CEREN

Ulrich S. Tran, University of Vienna
Martin Voracek, University of Vienna

Wolf Vanpaemel, University of Leuven
Aline Claesen, University of Leuven
Sara Gomes, University of Leuven
Thomas Verliefde, University of Leuven

Katherine Wick, Abilene Christian University
Ryan K. Jessup, Abilene Christian University
Monty L. Lynn, Abilene Christian University

Bradford J. Wiggins, Brigham Young University-Idaho
Scott D. Martin, Brigham Young University-Idaho
Samuel L. Clay, Brigham Young University-Idaho

Action Editor

Daniel J. Simons served as action editor for this article.

Author Contributions

B. Verschuere and E. H. Meijer proposed the replication project reported in this article and were responsible for writing the manuscript. B. Verschuere, E. H. Meijer, A. Jim, K. Hoogesteyn, and R. Orthey were responsible for developing and gathering the necessary materials. All the lead authors were involved in designing the overall procedure for the combined project that included the study reported by McCarthy et al. (2018, this issue). Each author contributed by conducting the study in his or her respective lab and providing valuable input on the manuscript.

Acknowledgments

The first two authors share first authorship. We thank Nina Mazar, On Amir, and Dan Ariely for providing materials for the study and for providing guidance about other tasks to include in the task battery; Chris Chabris for providing the abstract-reasoning task included as part of the battery; and Katherine Wood for assisting with the R scripts.

Declaration of Conflicting Interests

The author(s) declared that there were no conflicts of interest with respect to the authorship or the publication of this article.

Funding

This research was funded by Netherlands Organisation for Scientific Research (NWO) Grant 401.16.001/3873. The Association for Psychological Science and the Arnold Foundation provided funding to participating laboratories to defray the costs of running the study.

Supplemental Material

Additional supporting information can be found at http://journals.sagepub.com/doi/10.1177/2515245918781032

Open Practices

All data, analysis scripts, and materials have been made publicly available via the Open Science Framework. The data and scripts can be accessed at https://osf.io/mcvt7/wiki/home/, and the materials can be accessed at https://osf.io/rbejp/wiki/home/. The design and analysis plans were preregistered at the Open Science Framework and can be accessed at https://osf.io/3bwx5 and https://osf.io/hrju6/wiki/home/. The complete Open Practices Disclosure for this article can be found at http://journals.sagepub.com/doi/suppl/10.1177/2515245918781032. This article has received badges for Open Data, Open Materials, and Preregistration. More information about the Open Practices badges can be found at http://www.psychologicalscience.org/publications/badges.

Notes

1. Four labs were approved but did not meet the requirements of the protocol. Two labs (Huntjens, Sumampouw) changed aspects of the procedure that were important to the design, and two labs (Batra, Willis) recruited fewer than 200 participants. The data from the Huntjens lab are available on the OSF project page. The exclusion criteria specified in the protocol turned out to be vaguely worded, a problem we discovered as labs began to code their data. Consequently, we made a results-blind decision that the ancillary analyses of all labs contributing data would include labs that tested at least 200 participants before exclusions but fewer than 200 after exclusions. Data from these labs were excluded from the primary analysis and from the ancillary analyses of data from labs that strictly adhered to all protocol requirements.
2. This additional instruction went unnoticed by nine labs during the translation process (see Table 2). These nine labs were excluded from the ancillary analyses on data from labs that strictly met all inclusion criteria.
3. Although there was no heterogeneity at the lab level, there might still have been moderation of the effect at the individual level. Further analysis including moderators coded at the individual level might yield meaningful information about who cheats and who does not.

References

Ashton, M. C., & Lee, K. (2009). The HEXACO-60: A short measure of the major dimensions of personality. *Journal of Personality Assessment*, *91*, 340–345. doi:10.1080/00223890902935878

Ayal, S., Gino, F., Barkan, R., & Ariely, D. (2015). Three principles to REVISE people's unethical behavior. *Perspectives on Psychological Science*, *10*, 738–741. doi:10.1177/1745691615598512

Borenstein, M., Hedges, L. V., Higgins, J. P., & Rothstein, H. R. (2009). *Introduction to meta-analysis*. Chichester, England: John Wiley & Sons.

Gneezy, U. (2005). Deception: The role of consequences. *American Economic Review*, *95*, 384–394.

Guilford, J. P. (1967). *The nature of human intelligence*. New York, NY: McGraw-Hill.

Halevy, R., Shalvi, S., & Verschuere, B. (2014). Being honest about dishonesty: Correlating self-reports and actual lying. *Human Communication Research*, *40*, 54–72. doi:10.1111/hcre.12019

Mazar, N., Amir, O., & Ariely, D. (2008). The dishonesty of honest people: A theory of self-concept maintenance. *Journal of Marketing Research*, *45*, 633–644. doi:10.1509/jmkr.45.6.633

McCarthy, R. J., Skowronski, J. J., Verschuere, B., Meijer, E. H., Jim, A., Hoogesteyn, K., . . . Yıldız, E. (2018). Registered Replication Report on Srull and Wyer (1979). *Advances in Methods and Practices in Psychological Science*, *1*, 321–336.

McNair, D. M., Lorr, M., & Droppleman, L. F. (1971). *Profile of Mood States manual*. San Diego, CA: Educational and Industrial Testing Service.

Peer, E., Acquisti, A., & Shalvi, S. (2014). "I cheated, but only a little": Partial confessions to unethical behavior. *Journal of Personality and Social Psychology*, *106*, 202–217. doi:10.1037/a0035392

Pew Research Center. (2015). *America's changing religious landscape*. Retrieved from http://www.pewforum.org/2015/05/12/americas-changing-religious-landscape/

R Core Team. (2013). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing.

Rosenbaum, S. M., Billinger, S., & Stieglitz, N. (2014). Let's be honest: A review of experimental evidence of honesty and truth-telling. *Journal of Economic Psychology*, *45*, 181–196. doi:10.1016/j.joep.2014.10.002

Simons, D., Holcombe, A. O., & Spellman, B. A. (2014). An introduction to Registered Replication Reports at *Perspectives on Psychological Science*. *Perspectives on Psychological Science*, *9*, 552–555. doi:10.1177/1745691614543974

Slemrod, J. (2007). Cheating ourselves: The economics of tax evasion. *The Journal of Economic Perspectives*, *21*, 25–48. doi:10.1257/jep.21.1.25

Srull, T. K., & Wyer, R. S. (1979). The role of category accessibility in the interpretation of information about persons: Some determinants and implications. *Journal of Personality and Social Psychology*, *37*, 1660–1672. doi:10.1037/0022-3514.37.10.1660

The Tax Justice Network. (2011). *The cost of tax abuse: A briefing paper on the cost of tax evasion worldwide*. Retrieved from https://www.taxjustice.net/wp-content/uploads/2014/04/Cost-of-Tax-Abuse-TJN-2011.pdf

Thompson, S. G., & Higgins, J. P. (2002). How should meta-regression analyses be undertaken and interpreted? *Statistics in Medicine*, *21*, 1559–1573. doi:10.1002/sim.1187

Van Bavel, J. J., Mende-Siedlecki, P. M., Brady, W. J., & Reinero, D. A. (2016). Contextual sensitivity in scientific reproducibility. *Proceedings of the National Academy of Sciences, USA*, *113*, 6454–6459. doi:10.1073/pnas.1521897113

van Elk, M., Matzke, D., Gronau, Q. F., Guan, M., Vandekerckhove, J., & Wagenmakers, E.-J. (2015). Meta-analyses are no substitute for registered replications: A skeptical perspective on religious priming. *Frontiers in Psychology*, *6*, Article 1365. doi:10.3389/fpsyg.2015.01365